Agenda

Introduction to ESnet

How ESnet delivers its mission

Looking beyond the horizon
President’s National Objectives for DOE
Energy to Secure America’s Future

Quickly Implement the Economic Recovery Package: Create Millions of New Green Jobs and Lay the Foundation for the Future

Restore Science Leadership: Strengthen America’s Role as the World Leader in Science and Technology

Reduce GHG emissions: Drive emissions 20 Percent below 1990 levels by 2020

First Principle:
Pursue material and cost-effective measures with a sense of urgency

Enhance energy security: Save More Oil than the U.S Currently Imports from the Middle East and Venezuela Combined within 10 years

Enhance Nuclear Security: Strengthen non-proliferation activities, reduce global stockpiles of nuclear weapons, and maintain safety and reliability of the US stockpile
President’s National Objectives for DOE
Energy to Secure America’s Future

DOE’s Strategic Framework: Science and Discovery at the Core

ESnet exists solely to enable DOE’s science and discovery
• Single facility linking all 6 disciplines with their global collaborators

**ESnet Mission**
Provide DOE with interoperable, effective, and reliable communications infrastructure and leading-edge network services in support of the agency's missions
The Energy Sciences Network (ESnet)
A Department of Energy Facility

- National Fiber footprint
- Tier1 ISP
- Science Data Network
- International Collaborations
- Multiple 10G waves

Distributed Team of 35

Inder Monga ESnet
Network Planning Process

1) Exploring the plans and processes of the major stakeholders (the Office of Science programs, scientists, collaborators, and facilities):

1a) Data characteristics of scientific instruments and facilities
   - What data will be generated by instruments and supercomputers coming on-line over the next 5-10 years?

1b) Examining the future process of science
   - How and where will the new data be analyzed and used – that is, how will the process of doing science change over 5-10 years?

2) Understand all the Internet needs of DOE lab sites
   - Enterprise traffic profile (Web, Video, Email, SaaS…)

3) Observing current and historical network traffic patterns
   • What do the trends in network patterns predict for future network needs?
Science Process Evolution: From Vials to Visualization

- Instruments and Experiments
- Verifiable Results
- Visualization
- Collaboration and Sharing
- Large Simulations
- Distributed Data gathering and analysis

ESnet
Computational and Data Intensive Science

Scientific data growing exponentially

• Simulation systems and observational devices growing in capability exponentially
• Data sets may be large, complex, disperse, incomplete, and imperfect

Petabyte (PB) data sets common:

• *Climate modeling:* estimates of the next IPCC data is in 10s of petabytes
• *Genomics:* JGI alone will have ~1 petabyte of data this year and double each year
• *Particle physics:* LHC is projected to produce 16 petabytes of data per year
Large Science Requirements

**Bandwidth** – 100+ Gb/s core by 2012, 1TB in few links by 2015

**Reliability** – 99.999% availability for large data centers
  - Large instruments depend on the network, 24 x 7, to accomplish their science

**Global Connectivity - worldwide**
  - Geographic reach sufficient to connect users and analysis systems to Science facilities

**Services**
  - **Commodity IP is no longer adequate – guarantees are needed**
    - Guaranteed bandwidth, traffic isolation, service delivery architecture compatible with Web Services / Grid / “Systems of Systems” application development paradigms
    - Implicit requirement is that the service not have to pass through site firewalls which cannot handle the required bandwidth (frequently 10Gb/s)
  - **Visibility into the network end-to-end**
  - **Science-driven authentication infrastructure (PKI)**

**Assist users in effectively use the network**
  - Performance is always an application’s perspective
Planning for Growth - 80% YOY Growth

ESnet Accepted Traffic (TB/mo) - Log Scale

Log Plot of ESnet Monthly Accepted Traffic, January 1990 – Oct 2010

ESnet Traffic Increases by 10X Every 47 Months, on Average

Projected volume for Nov, 2011: 16629 TB
Actual volume for Nov, 2010: 10649 TB

Aug 1990 100 GBy/mo
Oct 1993 1 TBy/mo
Jul 1998 10 TBy/mo
Nov 2001 1 PBy/mo
Apr 2006 1 PBy/mo
Nov 2010 10 PBy/mo
Small Number of Large Flows Dominate

Red bars = top 1000 site to site workflows
Starting in mid-2005 a small number of large data flows dominate the network traffic
Note: as the fraction of large flows increases, the overall traffic increases become more erratic – it tracks the large flows

Orange bars = Virtual circuit flows

Overall ESnet traffic tracks the very large science use of the network
Keep It Simple but Smart principles

- Managing a nationwide network with < 20 engineers
- Automation and tools a huge part of the requirements
- Provisioning, troubleshooting, monitoring, customer support etc.
- Network Engineer + Software Developer combos!
Network Services for Science

High-Speed Data Transfer

- Hybrid Architecture
- **ESnet4**
- *Advanced Network Initiative (100G)*

Solving the end-to-end problem

- **Fasterdata**
  - http://fasterdata.es.net

Automated Network Resource Management

- Dynamic, Virtual Circuits for Science
  - **OSCARS**
    - http://www.es.net/oscars/

- Distributed network monitoring and troubleshooting
  - **perfSONAR**
    - http://perfsonar.net
Hybrid Architecture for ESnet4

The IP and SDN networks are fully interconnected and the link-by-link usage management implemented by OSCARS is used to provide a policy based sharing of each network by the other in case of failures.
On-Demand Secure Circuit and Advanced Reservation System (OSCARS)

• Original design goals
  – User requested bandwidth between specified points for a specific period of time
    • User request is via Web Services or a Web browser interface
    • Provide traffic isolation
  – Provide the network operators with a flexible mechanism for traffic engineering
    • E.g. controlling how the large science data flows use the available network capacity

• Learning through customer’s experience:
  – Flexible service semantics
    • E.g. allow a user to exceed the requested bandwidth, if the path has idle capacity – even if that capacity is committed (now)
  – Rich service semantics
    • E.g. provide for several variants of requesting a circuit with a backup, the most stringent of which is a guaranteed backup circuit on a physically diverse path (2011)

• Support the inherently multi-domain environment of large-scale science
  – Interoperate with similar services other network domains in order to set up cross-domain, end-to-end virtual circuits
Environment of Science is Inherently Multi-Domain

Inter-domain interoperability is crucial to serving science

In order to set up end-to-end circuits across multiple domains:

1. The domains exchange topology information containing at least potential VC ingress and egress points
2. VC setup request (via IDC protocol) is initiated at one end of the circuit and passed from domain to domain as the VC segments are authorized and reserved

Example – not all of the domains shown support the VC service
Network Mechanisms Underlying ESnet’s OSCARS

Layer 3 VC Service: Packets matching reservation profile IP flow-spec are filtered out (i.e. policy based routing), “policed” to reserved bandwidth, and injected into an LSP.

Layer 2 VC Service: Packets matching reservation profile VLAN ID are filtered out (i.e. L2VPN), “policed” to reserved bandwidth, and injected into an LSP.

Best-effort IP traffic can use SDN, but under normal circumstances it does not because the OSPF cost of SDN is very high.
OSCARS is a Production Service in ESnet

OSCARS is currently being used to support production traffic ≈ 50% of all ESnet traffic is now carried in OSCARS VCs

Operational Virtual Circuit (VC) support

- As of 11/2010, there are ~33 (up from 26 in 10/2009) long-term production VCs instantiated
  - 25 VCs supporting HEP: LHC T0-T1 (Primary and Backup) and LHC T1-T2
  - 3 VCs supporting Climate: NOAA Global Fluid Dynamics Lab and Earth System Grid
  - 2 VCs supporting Computational Astrophysics: OptiPortal
  - 1 VC supporting Biological and Environmental Research: Genomics

- Short-term dynamic VCs
  - Between 1/2008 and 6/2010, there were roughly 5000 successful VC reservations
    - 3000 reservations initiated by BNL using TeraPaths
    - 900 reservations initiated by FNAL using LambdaStation
    - 700 reservations initiated using Phoebus
    - 400 demos and testing (SC, GLIF, interoperability testing (DICE))

Helped ESnet in winning Excellence.gov “Excellence in Leveraging Technology” award and InformationWeek’s 2009 “Top 10 Government Innovators” Award

*A TCP path conditioning approach to latency hiding - [http://damsl.cis.udel.edu/project/phoebus/](http://damsl.cis.udel.edu/project/phoebus/)
OSCARS Open-Source Software
(http://code.google.com/p/oscars-idc/)

The code base is undergoing its third rewrite (OSCARS v0.6)

• Make it more modular and expose internal APIs
  - For example, ability to plug and play your own PCE
  - Targeted to facilitate research collaborations

• As the service semantics get more complex (in response to user requirements) focus “complex, compound network services”
  - Defining “atomic” service functions and building mechanisms for users to \textit{compose} these building blocks into custom services
OSCARS Version 0.6 Software Architecture

- **Notification Broker**
  - Manage subscriptions
  - Forward notifications

- **AuthN**
  - Authentication

- **AuthZ**
  - Authorization
  - Costing
  - *Distinct data and control plane functions*

- **Web Browser User Interface**

- **Coordinator**
  - Workflow coordinator

- **Lookup Bridge**
  - Lookup service

- **Topology Bridge**
  - Topology information management

- **Path Computation Engine**
  - Constrained path computations

- **Path Setup**
  - Network element interface

- **Resource Manager**
  - Manage reservations
  - Auditing

- **Web Services API**
  - Manages external WS communications

- **SOAP + WSDL over http/https**

- **perfSONAR services**

- **The lookup and topology services are now seconded to perfSONAR**

- Other IDCs

- User apps

The lookup and topology services are now seconded to perfSONAR.
Why does the Network seem so slow?
Importance of end-to-end network performance for science

Very large files and very large flows
- 10s to 100s of GB
- Single flow rates of 100s to 1000s of Mbps
- Network latency from 10s of msec to over 300msec

Packet loss must be essentially zero
- Zero packet loss essential for multi-gigabit performance
- Latency and packet loss interact in very unpleasant ways
- Not true for commodity ISP / carrier networks

Large buffers are critical
- Data center and LAN switch platforms typically have tiny interface buffers
- When placed in the path of wide area data transfers, these devices cause performance problems
- Therefore, data-intensive science cannot simply be put onto commodity data center platforms
Where are common problems?

- Congested or faulty links between domains
- Latency dependant problems inside domains with small RTT
Local testing will not find all problems

Performance is poor when RTT exceeds 20 ms

Performance is good when RTT is < 20 ms

Switch with small buffers
Network performance measurements infrastructure: perfSONAR

- Multi-service, distributed infrastructure
- Distributed, rapid troubleshooting and fault isolation
- Latency and packet loss measurement
- Collaboration
- Deployed in a large number of research networks

http://weathermap.es.net
**perfSONAR Architecture**

<table>
<thead>
<tr>
<th>layer</th>
<th>architectural relationship</th>
<th>examples</th>
</tr>
</thead>
</table>
| interface           | human user ➔ performance GUI → client (e.g. part of an application system communication service manager) | • real-time end-to-end performance graph (e.g. bandwidth or packet loss vs. time)  
                     |                                                                               | • historical performance data for planning purposes                        
                     |                                                                               | • event subscription service                                                 |
| service             | path monitor ➔ topology aggregator ➔ measurement archive(s) ➔ event subscription service   |                                                                           |
|                     | service locator ➔ measurement export                                                       |                                                                           |
| measurement point   | measurement export ➔ measurement export                                                    | • The measurement points (m1….m6) are the real-time feeds from the network or local monitoring devices  
                     |                                                                           | • The Measurement Export service converts each local measurement to a standard format for that type of measurement |
ESnet widely deploys perfSONAR

perfSONAR nodes deployed next to all backbone routers, and at all 10Gb connected sites

- 31 locations deployed
- Full list of active services at:
  - http://www.perfsonar.net/activeServices/

Instructions on using these services for network troubleshooting:

- http://fasterdata.es.net

Federated information is extremely useful to help debug a number of problems

- The only tool that we have to monitor circuits end-to-end across the networks from the US to Europe

PerfSONAR measurement points are deployed at dozens of R&E institutions in the US and more in Europe

- See https://stats1.es.net/perfSONAR/directorySearch.html

The value of perfSONAR increases as it is deployed at more sites
Looking beyond the horizon

Lawrence Berkeley National Lab
A legacy of improving our lives and understanding the world around us

Discovered 16 elements
Identified good and bad cholesterol
Confirmed the Big Bang and discovered Dark Energy
Turned windows into energy savers
Unmasked a dinosaur killer
Exposed the Radon risk
Explained photosynthesis
Created the toughest ceramic
Pitted cool roofs against global warming
Given fluorescent lights their big break
Caught Malaria in the act
Built a better battery
Preserved the sounds of yesteryear
Fabricated the smallest machines
Made appliances pull their weight
Brought safe drinking water to thousands
Created a pocket-sized DNA sampler

• Revealed the secrets of the human genome
• Redefined the causes of breast cancer
• Given buildings an energy makeover
• Supercharged the climate model
• Derailed an ecological danger
• Helped bring energy efficiency to China
• Pioneered medical imaging
• Brought the stars closer
ARRA Advanced Networking Initiative (ANI)

ESnet received ~$62M in ARRA funds from DOE for an Advanced Networking Initiative to:

- Build an end-to-end 100 Gbps prototype network
- Handle proliferating data needs between the three DOE supercomputing facilities and NYC international exchange point
- Build a network testbed facility for researchers and industry

DOE is also funding $5M in network research using the testbed facility: goal of near-term technology transfer to the production ESnet network.
ARRA Magellan Initiative

Separately funded $33 million for Magellan, an associated DOE cloud computing project that will utilize the 100 Gbps network infrastructure

- Establish a nationwide scientific mid-range distributed computing and data analysis testbed
- Two sites (NERSC / LBNL and ALCF / ANL) planned
- Multiple 10’s of teraflops and multiple petabytes of storage, as well as appropriate cloud software tuned for moderate concurrency.
Prototype 100G Topology
Testbed Overview
An Open Facility

Progression:
• Start out as a tabletop testbed, then move out to the wide-area when 100 Gbps available

Capabilities:
• Ability to support end-to-end networking, middleware and application experiments, including interoperability testing of multi-vendor 100 Gbps network components
• Dynamic network provisioning
• Plan to acquire dark fiber on a portion of testbed footprint to enable hybrid (layer 0-3) network research
• Use Virtual Machine technology to support protocol and middleware research
• Detailed monitoring so researchers will have access to all possible monitoring data from the network devices
Tabletop: A layered view

- Layer 0/1
- Layer 2/Openflow
- Layer 3

- IO Tester
- App host
- Monitoring Host
- Compute/Storage
- IO Testers
- WDM Link
- 10GE Link
- 1GE Link

VMs

Research Applications

Lawrence Berkeley National Laboratory

U.S. Department of Energy | Office of Science

Inder Monga ESnet
Sample Configuration: Multi-Domain Multi-Layer Protection Testing

North Domain

South Domain

Test inter-domain optical protection schemes

East Domain

Test inter-domain higher layer (> 1) protection schemes

Openflow Switch

IO Tester

IOTester

north-wdm1

north-wdm2

east-wdm1

east-wdm2

South Domain

Test inter-domain optical protection schemes

north-wdm1

south-wdm1

south-wdm2

Openflow Switch

IO Tester

RTR

IOTester

RTR
ESnet Research
Solving hard problems collaboratively
Atomic and Composite Network Services Architecture

Network Service Plane

Composite Service (S1 = S2 + S3)

Composite Service (S2 = AS1 + AS2)

Composite Service (S3 = AS3 + AS4)

Atomic Service (AS1)

Atomic Service (AS2)

Atomic Service (AS3)

Atomic Service (AS4)

Network Services Interface

e.g. a backup circuit—be able to move a certain amount of data in or by a certain time

e.g. monitor data sent and/or potential to send data

Service template pre-composed for specific applications or customized by advanced users

Atomic services used as building blocks for composite services

Multi-Layer Network Data Plane

e.g. dynamically manage priority and allocated bandwidth to ensure deadline completion

Service Abstraction Increases Service Usage Simplifies

Network Services Interface

Service Abstraction Increases Service Usage Simplifies

Network Services Interface

Service Abstraction Increases Service Usage Simplifies
Examples of Composite Network Services

LHC: Resilient High Bandwidth Guaranteed Connection
- connect
- topology
- find path
- protect
- measure
- monitor

Reduced RTT Transfers: Store and Forward Connection

Protocol Testing: Constrained Path Connection
Atomic Network Services Currently Offered by OSCARS

ESnet OSCARS

**Connection** creates virtual circuits (VCs) within a domain as well as multi-domain end-to-end VCs

**Path Finding** determines a viable path based on time and bandwidth constrains

**Monitoring** provides critical VCs with production level support

Multi-Layer Multi-Layer Network Data Plane
Multi-layer networking

Virtual Organization Specific View

Friday, 3am Maintenance View

IP View

What paths are possible between A and B, between time i and ii, with specific policy.

User Specified Multiple Views

ESnet IP

ESnet SDN

WDM Layer

OSPF-TE

Topology

SDN View

• Time Domain
• AAA
• Management (SNMP) Data
• Administrator Requirements

MX-TCE
Thank You!

Contact: imonga [at] es.net